



AFRL-RI-RS-TR-2013-026

## **BIOLOGICALLY INSPIRED CIRCUITS FOR VISUAL SEARCH AND RECOGNITION IN COMPLEX SCENES**

---

MASSACHUSETTS INSTITUTE OF TECHNOLOGY /  
HARVARD UNIVERSITY

*FEBRUARY 2013*

FINAL TECHNICAL REPORT

***APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED***

STINFO COPY

**AIR FORCE RESEARCH LABORATORY  
INFORMATION DIRECTORATE**

## NOTICE AND SIGNATURE PAGE

Using Government drawings, specifications, or other data included in this document for any purpose other than Government procurement does not in any way obligate the U.S. Government. The fact that the Government formulated or supplied the drawings, specifications, or other data does not license the holder or any other person or corporation; or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

This report is the result of contracted fundamental research deemed exempt from public affairs security and policy review in accordance with SAF/AQR memorandum dated 10 Dec 08 and AFRL/CA policy clarification memorandum dated 16 Jan 09. This report is available to the general public, including foreign nationals. Copies may be obtained from the Defense Technical Information Center (DTIC) (<http://www.dtic.mil>).

AFRL-RI-RS-TR-2013-026 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

FOR THE DIRECTOR:

/ S /

YURIY LUZANOV  
Work Unit Manager

/ S /

WARREN H. DEBANY, JR.  
Technical Advisor, Information Exploitation  
and Operations Division  
Information Directorate

This report is published in the interest of scientific and technical information exchange, and its publication does not constitute the Government's approval or disapproval of its ideas or findings.

# REPORT DOCUMENTATION PAGE

Form Approved  
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Washington Headquarters Service, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188) Washington, DC 20503.

**PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.**

1. REPORT DATE (DD-MM-YYYY) FEBRUARY 2013	2. REPORT TYPE FINAL TECHNICAL REPORT	3. DATES COVERED (From - To) OCT 2010 – OCT 2012		
4. TITLE AND SUBTITLE  BIOLOGICALLY INSPIRED CIRCUITS FOR VISUAL SEARCH AND RECOGNITION IN COMPLEX SCENES		5a. CONTRACT NUMBER FA8750-11-2-0009		
		5b. GRANT NUMBER N/A		
		5c. PROGRAM ELEMENT NUMBER 62788F		
6. AUTHOR(S) Tomaso Poggio (Massachusetts Institute of Technology) and Gabriel Kreiman (Harvard University)		5d. PROJECT NUMBER E2AS		
		5e. TASK NUMBER PR		
		5f. WORK UNIT NUMBER TY		
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Massachusetts Institute of Technology 43 Vassar St, Bldg 46-5155 Cambridge, MA 02139		8. PERFORMING ORGANIZATION REPORT NUMBER N/A		
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) Air Force Research Laboratory/RIGC 525 Brooks Road Rome NY 13441-4505		10. SPONSOR/MONITOR'S ACRONYM(S) N/A		
		11. SPONSORING/MONITORING AGENCY REPORT NUMBER AFRL-RI-RS-TR-2013-026		
12. DISTRIBUTION AVAILABILITY STATEMENT Approved for Public Release; Distribution Unlimited. This report is the result of contracted fundamental research deemed exempt from public affairs security and policy review in accordance with SAF/AQR memorandum dated 10 Dec 08 and AFRL/CA policy clarification memorandum dated 16 Jan 09.				
13. SUPPLEMENTARY NOTES				
14. ABSTRACT Here we report on the advances and insights derived from studying and developing computational algorithms for object recognition, feature-based attention and visual search. As part of these efforts, we have created and documented software that includes feed-forward combination of selectivity and tolerance in visual recognition feedback signals for attention, search and object completion. We have quantitatively characterized and evaluated the performance of the system under a variety of different recognition problems with varying levels of difficulty, different levels of approximation to real-world recognition problems and different degrees of temporal dynamics. These measurements provide state-of-the-art benchmarks for different recognition problems. In particular, we evaluated (a) Single objects on uniform backgrounds and transformations of those objects (scale, position, viewpoint, illumination); (b) Combination of multiple objects on uniform backgrounds; (c) Single objects embedded in natural backgrounds; (d) Faces and objects in commercial movies. We have made progress on three main fronts that involve extensions and improvements to the existing software: (i) addition of feedback and recognition of occluded objects; (ii) Initial optimization of radial basis function centers in intermediate processing stages; (iii) Visual search in cluttered scenes.				
15. SUBJECT TERMS Object recognition, visual system feedback, biologically inspired vision, ventral stream modeling				
16. SECURITY CLASSIFICATION OF:  a. REPORT U		17. LIMITATION OF ABSTRACT SAR	18. NUMBER OF PAGES 28	19a. NAME OF RESPONSIBLE PERSON YURIY LUZANOV
				19b. TELEPHONE NUMBER (Include area code) (315) 330-3489

## Table of Contents

1. Summary.....	1
2. Introduction.....	1
3. Methods, Assumptions and Procedures .....	3
3.1 Bottom-up architecture .....	3
3.2 Feed-forward and feed-back architecture.....	5
4. Results and Discussion.....	5
4.1 Object completion .....	5
4.2 Characterizing the model's performance and RBF optimization.....	11
4.3 Recognition during dynamic viewing conditions .....	13
4.4 Visual search .....	16
4.5 Improvements to prototypes.....	18
4.6 Publications.....	20
5. Conclusions.....	21
6. References.....	23

## **List of figures**

Figure 1: Bottom-up model schematic .....	2
Figure 2: Sketch of the FF/FB model.....	4
Figure 3. Top-down signals can improve object completion .....	6
Figure 4. Feed-forward / feed-back model for object completion .....	7
Figure 5: Performance of the model for partial inputs (example).....	8
Figure 6: Performance of the model for partial inputs (summary) .....	9
Figure 7: Example stimuli and transformations .....	10
Figure 8: Object transformations and model performance.....	11
Figure 9: Towards studying video sequences.....	12
Figure 10: Manual annotation scheme .....	13
Figure 11: Example semi-supervised video annotations .....	14
Figure 12: Example stimuli for visual search task .....	16
Figure 13: The model can search for target objects .....	17
Figure 14: A normalization operation is critical for search model.....	17
Figure 15: Comparison between model and human performance.....	18
Figure 16: Schematic proposal to study video sequences .....	19

## 1. Summary

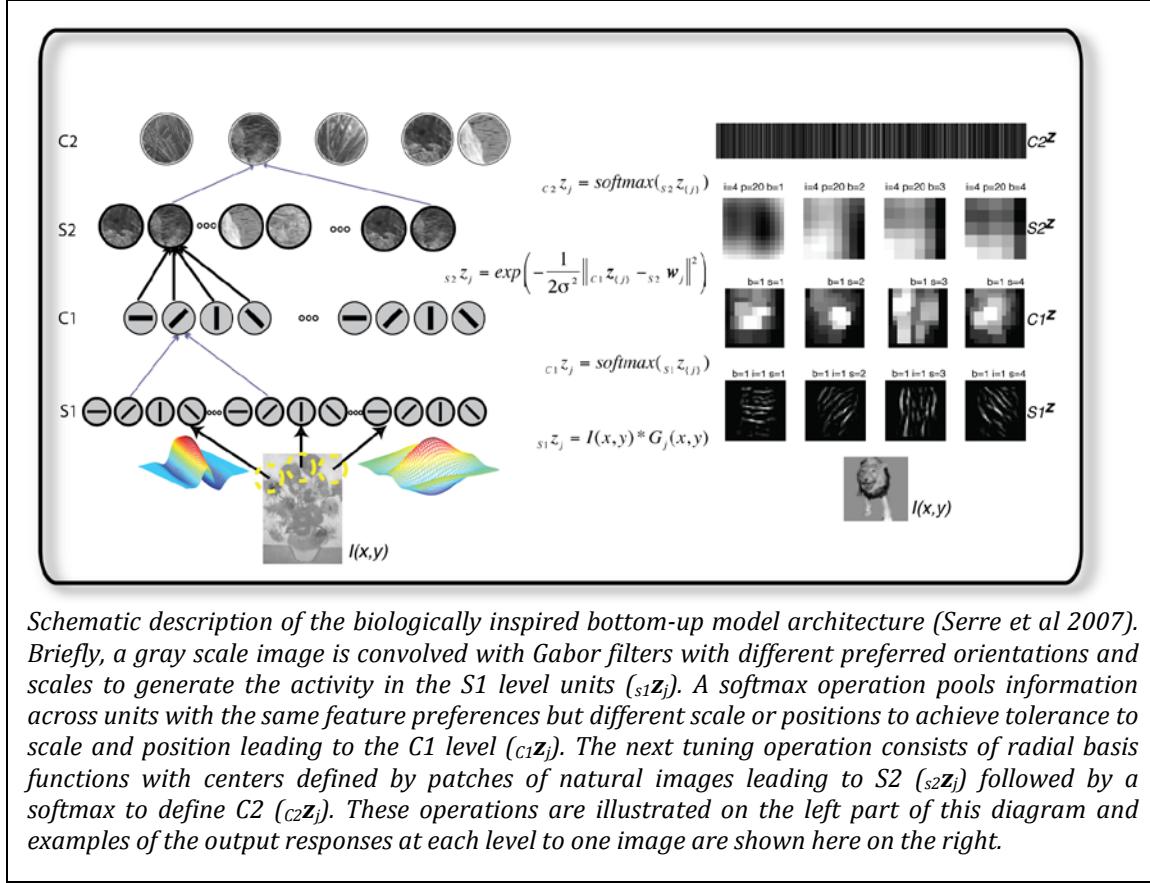
Here we report on the advances and insights derived from studying and developing computational algorithms for object recognition, feature-based attention and visual search. As part of these efforts, we have created and documented software that includes feed-forward combination of selectivity and tolerance in visual recognition feedback signals for attention, search and object completion. We have quantitatively characterized and evaluated the performance of the system under a variety of different recognition problems with varying levels of difficulty, different levels of approximation to real-world recognition problems and different degrees of temporal dynamics. These measurements provide state-of-the-art benchmarks for different recognition problems. In particular, we evaluated (a) Single objects on uniform backgrounds and transformations of those objects (scale, position, viewpoint, illumination); (b) Combination of multiple objects on uniform backgrounds; (c) Single objects embedded in natural backgrounds; (d) Faces and objects in commercial movies. We have made progress on three main fronts that involve extensions and improvements to the existing software: (i) addition of feedback and recognition of occluded objects; (ii) Initial optimization of radial basis function centers in intermediate processing stages; (iii) Visual search in cluttered scenes.

## 2. Introduction

Object recognition in cortex is mediated by the ventral visual pathway running from primary visual cortex (V1) (Hubel and Wiesel, 1959) through extrastriate visual areas V2 and V4 to inferior temporal cortex (ITC). Information from ITC is conveyed to prefrontal cortex (PFC) which is involved in linking perception to memory and action. Over the last decade, a number of physiological studies in non-human primates have established several basic facts about the cortical mechanisms of recognition. The accumulated evidence points to key features of the ventral pathway. From V1 to ITC, there is an increase in invariance to position and scale and, in parallel, an increase in the size of the receptive fields as well as in the complexity of the optimal stimuli for the neurons.

One of the first feedforward models, Fukushima's Neocognitron (Fukushima, 1980), followed the basic Hubel & Wiesel hierarchy in a computer vision system. Building upon several existing neurobiological models, conceptual proposals and computer vision systems, we have been developing a similar computational theory that attempts to quantitatively account for a host of recent anatomical and physiological data.

The connectivity between layers of the model is described in **Figure 1**. The connectivity in the system is inspired by neuroanatomical constraints (Felleman and Van Essen, 1991). An important aspect of this new model is a developmental-like,



**Figure 1: Bottom-up model schematic**

unsupervised learning stage in which each unit in the S layers become tuned to a different patch of a natural image (see (Serre et al., 2005b; Serre et al., 2007)). In this way, the model develops a generic dictionary of shape-components from V2 to ITC, which provides a rich representation for task-specific categorization circuits in higher brain areas. In addition to the S layers, the C layers pool information from S units to achieve tolerance to transformations of the input. Because target objects can be physically similar to distractor clutter and because objects can cast an infinite number of projections on the retina, this combination of selectivity and transformation tolerance is key to visual recognition.

The hierarchical architecture builds a representation that shows progressively more invariance to position and scale while preserving the selectivity of the units. The resulting dictionary provided by the S and C layers at different levels of the hierarchy is generic and universal, in the sense that it can support several different recognition tasks and, in particular, the recognition of many different object categories (Freedman et al., 2001; Serre et al., 2005b). For the “mature” model to learn a new categorization task, only the task-specific circuits at the top level in the model, possibly corresponding to categorization units in PFC, have to be trained from a small set of labeled examples and in a task specific manner.

The model attempts to be close to the anatomy and the physiology of visual

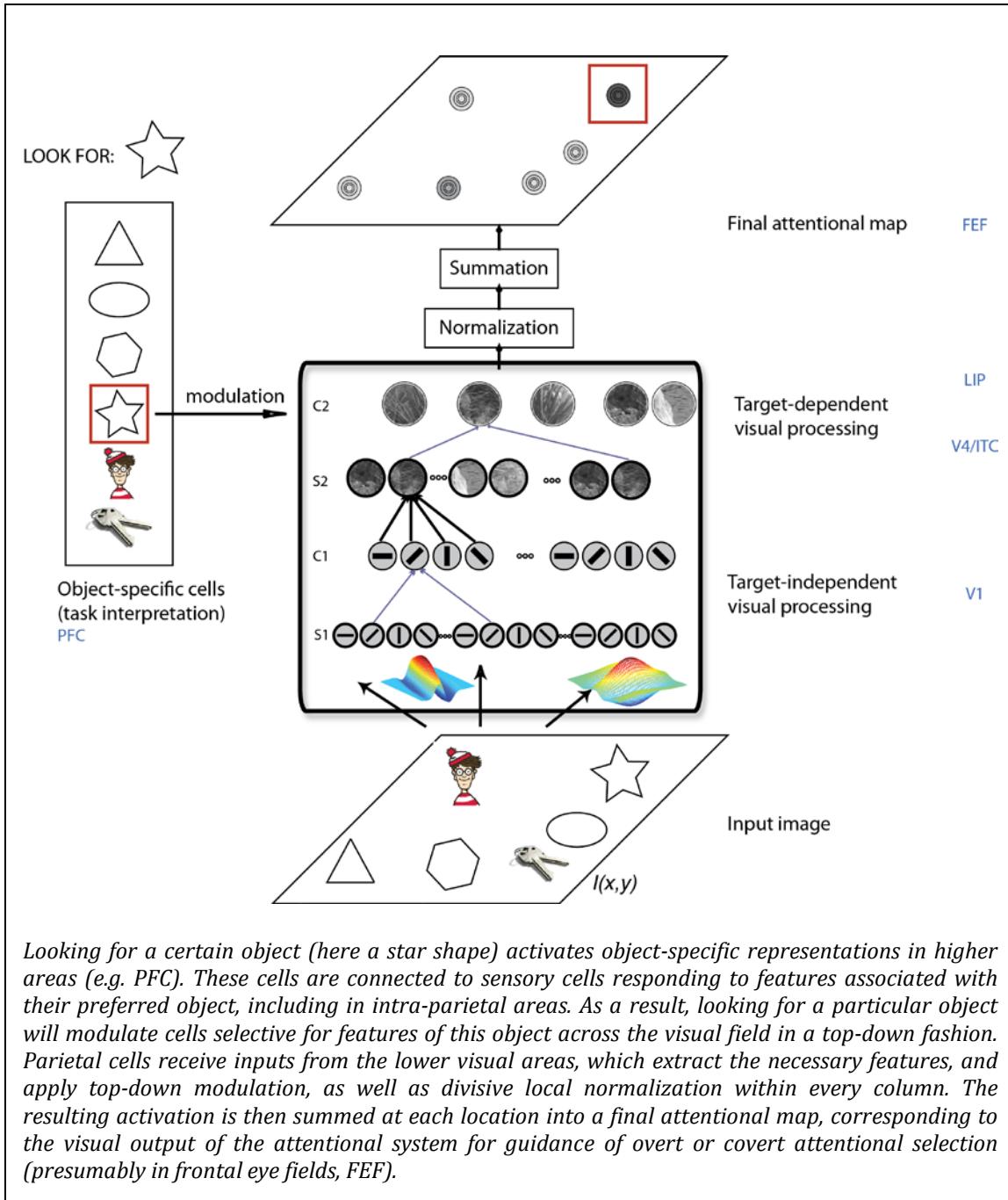
cortex in terms of quantitative parameter values. It is qualitatively and quantitatively consistent with (and in some cases actually predicts) several properties of cells in V1, V4, IT and PFC as well as functional imaging and psychophysical data. For instance, the model predicts the max-like behavior of a subclass of complex cells in V1 and V4. It also agrees with other data in V4 about the response of neurons to combinations of simple two-bar stimuli (within the receptive field of the S2 units) and some of the C2 units in the model show a tuning for boundary conformations which is consistent with recordings from V4. Read-out from C2b units in the model predicted recent read-out experiments in ITC (Hung et al., 2005), showing very similar selectivity and invariance for the same set of stimuli. The model performs at the level of the best computer vision systems in recognizing objects in the real world (Serre et al., 2005b; Serre et al., 2005a; Bileschi, 2006). The model is remarkably robust to parameter values, detailed wiring and exact form of the two basic operations and of the learning rule. The model also matches human performance – when eye movements and attentional effects are not allowed — in a difficult categorization task involving natural images of a class of objects in context and background.

While previous instantiations of the model involved purely feedforward (bottom-up) processing, our more recent developments include the addition of feedback (top-down) influences that can implement attentional modulation and Bayesian inference (Chikkerur et al., 2009; Chikkerur et al., In Press). From an anatomical standpoint, there are more backprojections in cortex than feedforward projections. Our computational algorithm combines bottom-up and top-down signals enhancing search and recognition in complex scenes. This biologically inspired cortex-like circuit constrained by the neuroanatomy and neurophysiology of the ventral visual cortex as well as by psychophysical investigations of human performance in recognition tasks, constitutes the key element in this proposal.

Two important points of interest are the ability to detect small objects in cluttered scenes and to be able to scale up the model to large datasets. Most of the video that Air Force collects from aerial platforms has a lot of clutter. This is especially the case with wide area surveillance platforms. It is important to be able to identify individual objects in this type of data. The biologically inspired algorithms incorporating both bottom-up and top-down signals are ideally suited for this task. Eventually, in an automated system, it will be necessary to discriminate between many different targets. Therefore it is important to study the scalability of this approach. The other side of scalability is the necessary computational requirements to achieve near real-time performance. One particular set of resources that can be utilized for this purpose are GPGPUs (Mutch et al., 2010).

### **3. Methods, Assumptions and Procedures**

#### **3.1 Bottom-up architecture**



Looking for a certain object (here a star shape) activates object-specific representations in higher areas (e.g. PFC). These cells are connected to sensory cells responding to features associated with their preferred object, including in intra-parietal areas. As a result, looking for a particular object will modulate cells selective for features of this object across the visual field in a top-down fashion. Parietal cells receive inputs from the lower visual areas, which extract the necessary features, and apply top-down modulation, as well as divisive local normalization within every column. The resulting activation is then summed at each location into a final attentional map, corresponding to the visual output of the attentional system for guidance of overt or covert attentional selection (presumably in frontal eye fields, FEF).

**Figure 2: Sketch of the FF/FB model**

A schematic illustration of the CBCL model is shown in **Figure 1**. This instantiation of the model comprises purely bottom-up signals and has been described in previous publications (Riesenhuber and Poggio, 1999; Serre et al., 2007). Briefly, the system has the following properties: (1) bottom-up architecture; (2) increase in receptive field size along the hierarchy; (3) increase in the complexity of encoded features along the hierarchy as a consequence of a tuning operation; (4) increase in tolerance to feature transformations as a consequence of

non-linear max-like operation; (5) intercalation of tuning / tolerance operations. Code implementing this system (including a GPU-based framework and multiple different versions) can be found at <http://cbcl.mit.edu/software-datasets/index.html>

### 3.2 Feed-forward and feed-back architecture

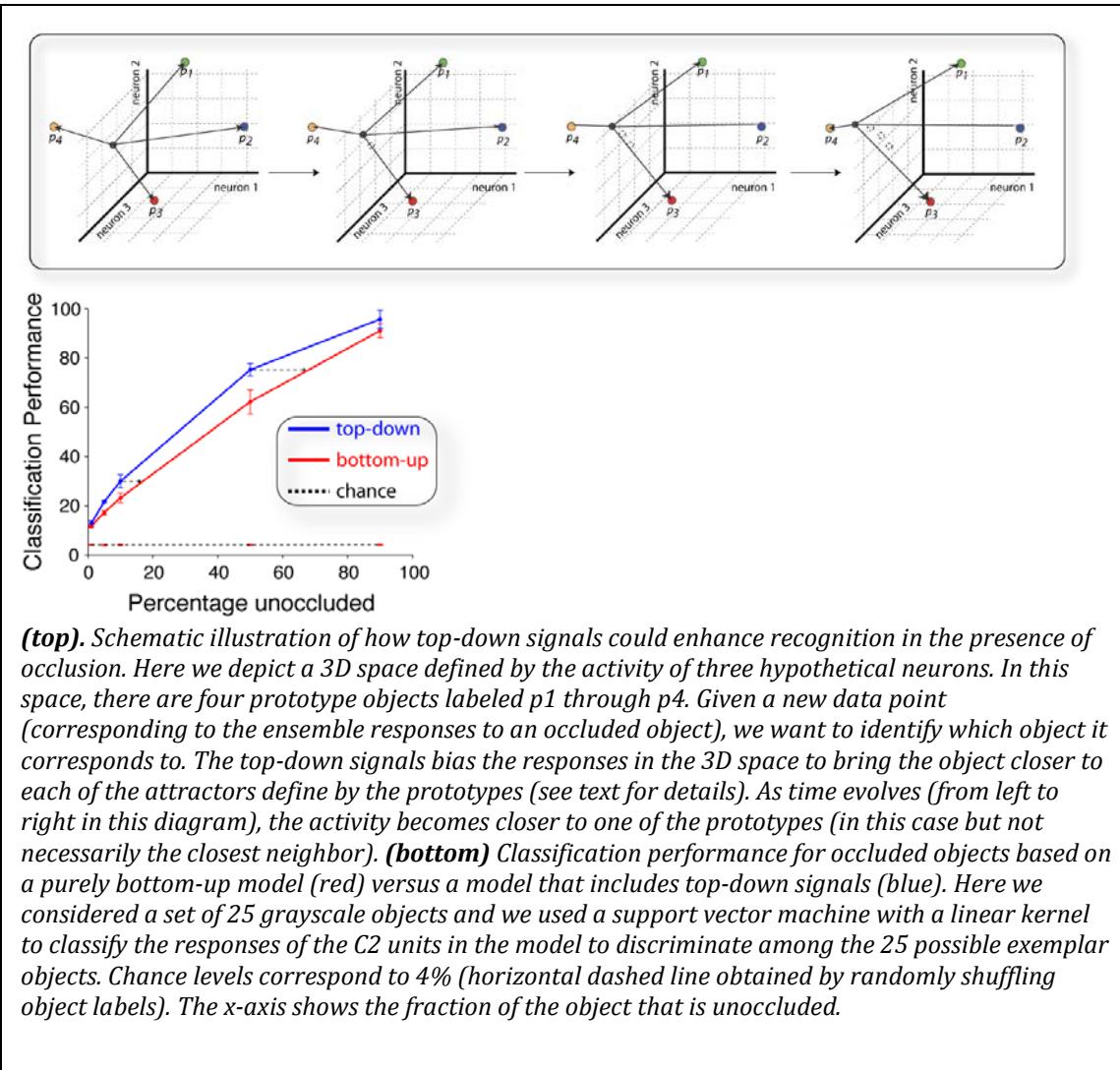
A number of previous models have incorporated feed-back signals onto the architecture shown in **Figure 1** (or similar architectures) in order to describe the role of attentional signals in enhancing specific locations within the stimulus. A demonstration of the power of combining spatial attention and feed-forward signals to locate objects and identify them (“What is where”) has been documented in the work of Chikkerur (Chikkerur et al., 2010). We have extended this work now include “feature attention”. Feature attention refers to the possibility of enhancing specific aspects of a stimulus (e.g. color, shape, motion) rather than a specific location. A typical scenario involves searching for a face in a crowd or searching for a red element in a crowded display. A schematic of the model including feed-forward and feedback connections is shown in **Figure 2**. The performance of this model in different search tasks and the comparison with human performance is described below.

## 4. Results and Discussion

### 4.1 Object completion

Clutter is an important problem in visual recognition because objects are typically embedded in complex natural scenes (and often camouflaged) and this significantly impairs the performance of many automatic recognition algorithms. Of note, the activity of neurons along the primate ventral visual pathways is also significantly reduced in the presence of clutter. Thus, a central aspect of this proposal is to characterize and solve the problem of clutter in visual recognition by combining the power of bottom-up processing (model depicted in **Figure 1**) with efficient top-down signals that can ameliorate or even eliminate the problem of clutter.

An extreme version of clutter arises when objects overlap in the retinal projection so that the object of interest is partially occluded. In these cases, the recognition machinery needs to be able to reconstruct the object from partial information; the process is usually referred to as “object completion”. Of course, in the extreme of minimal or no occlusion, the problem of recognition reduces to the situation of object recognition in clutter and in the other extreme of 100% occlusion, the problem becomes impossible. In between these two extremes, there is an interesting and important regime where we may be able to recognize partially occluded objects.



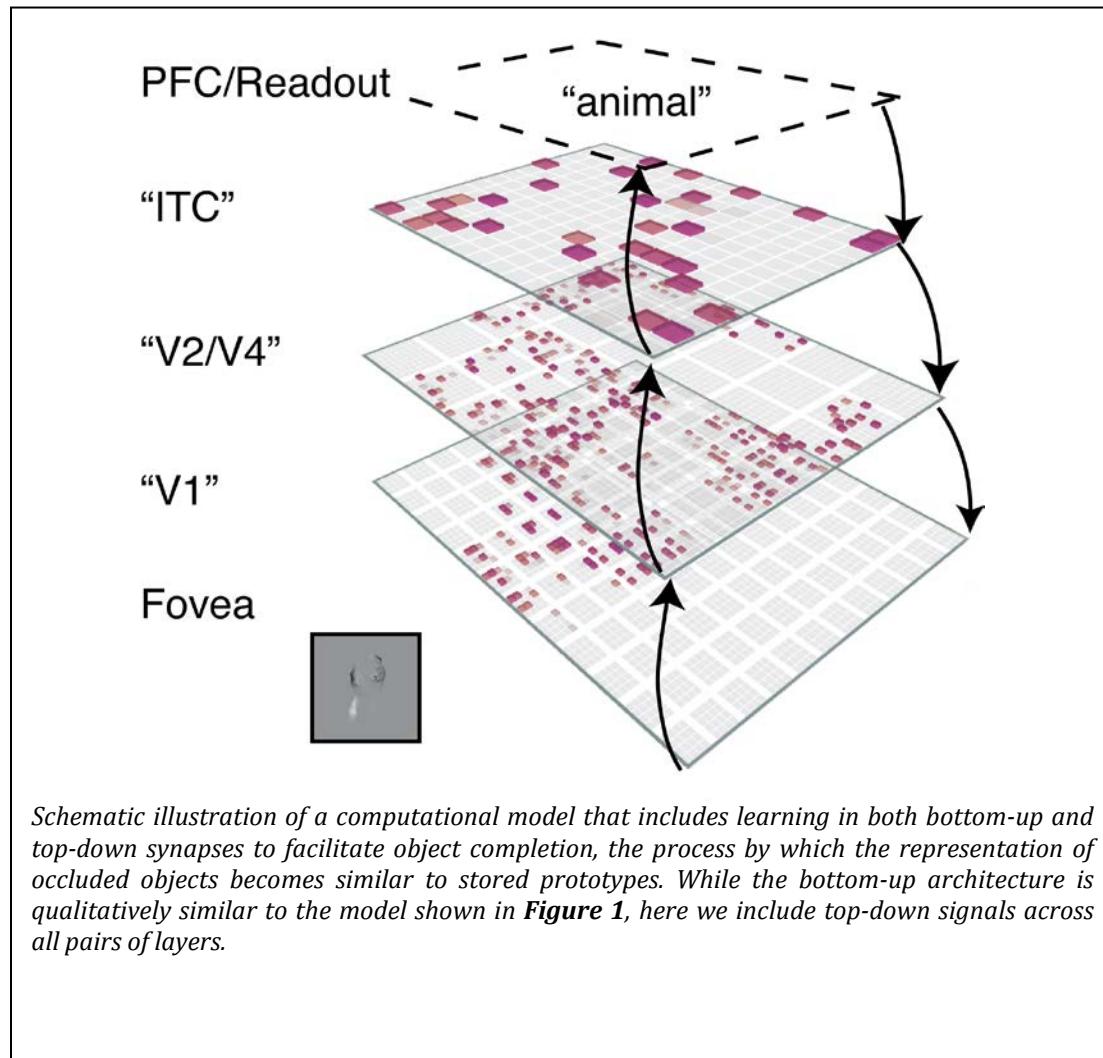
**Figure 3. Top-down signals can improve object completion**

We conjecture that object completion from partial information represents another instance where top-down signals will play a critical role. Essentially, there is a limitless number of possibilities to reconstruct an object; to constrain the problem, it is necessary to use prior information based on knowledge about the objects. We can think of this as part of the priors in the Bayesian approach formulated by Chikkerur et al (2010). **Figure 3** illustrate our preliminary efforts to perform object completion using top-down signals added to the bottom-up architecture used in Serre et al 2007. The basic intuition is to compare the bottom-up responses to the stored memories of specific prototype images and then use top-down signals to bias the activity in earlier stages towards those prototypes. The procedure is illustrated in a schematic fashion in **Figure 3 (top)** in a situation that includes a 3-dimensional space defined by three neurons and a task consisting of identifying an object among 4 possible prototypes. The colored circles denoted  $p_1$  through  $p_4$  show the position of the prototypes in the 3-d space. Given a new data

point (corresponding to the activity of this ensemble of neurons in response to an occluded object), which object does it belong to? The top-down signals act as attractors towards the four possible prototypes according

to:  $c_2 z(t) = c_2 z(t-1) + \sum_i \frac{\alpha(p_i - c_2 z(t-1))}{d(z(t-1), p_i)^n}$ . Here  $c_2 Z(t)$  represents the activity at the

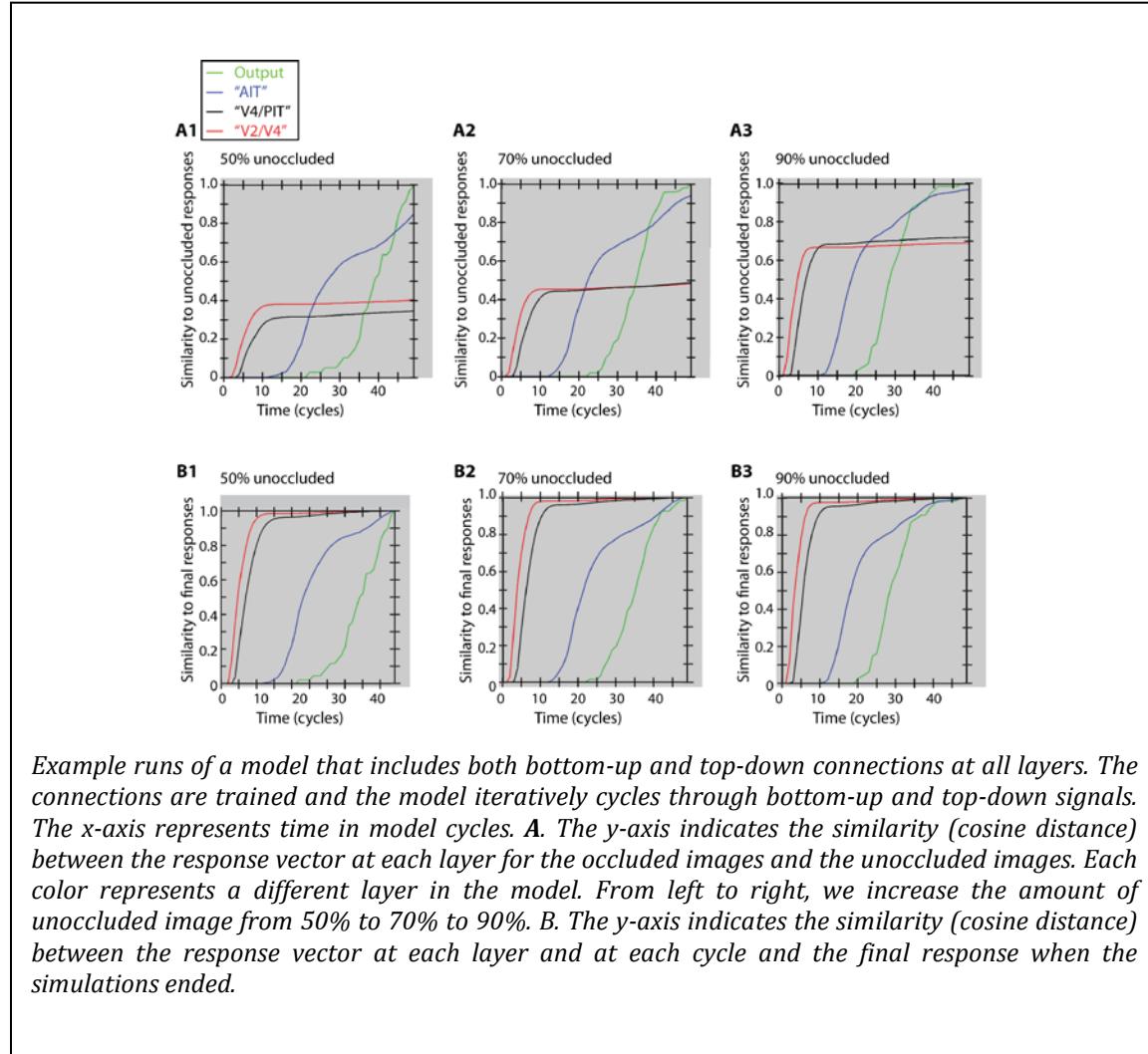
“C2” level in **Figure 1**,  $t$  indicates time,  $p_i$  indicates the position in C2 space for prototype number  $i$ ,  $d$  is the Euclidian distance between two points in C2 space and  $\alpha, n$  are two tuning parameters that control the relative weight of each attractor and how the attractor strength decreases with distance.



**Figure 4. Feed-forward / feed-back model for object completion**

As a proof of principle, we considered a situation with 25 different exemplar objects and we used a technique called “Bubbles” to systematically introduce occlusion (Gosselin and Schyns, 2001). The results of our preliminary analyses are

shown in **Figure 3 (bottom)**. The y-axis indicates the classification performance based on a support vector machine with a linear kernel trained to learn the map

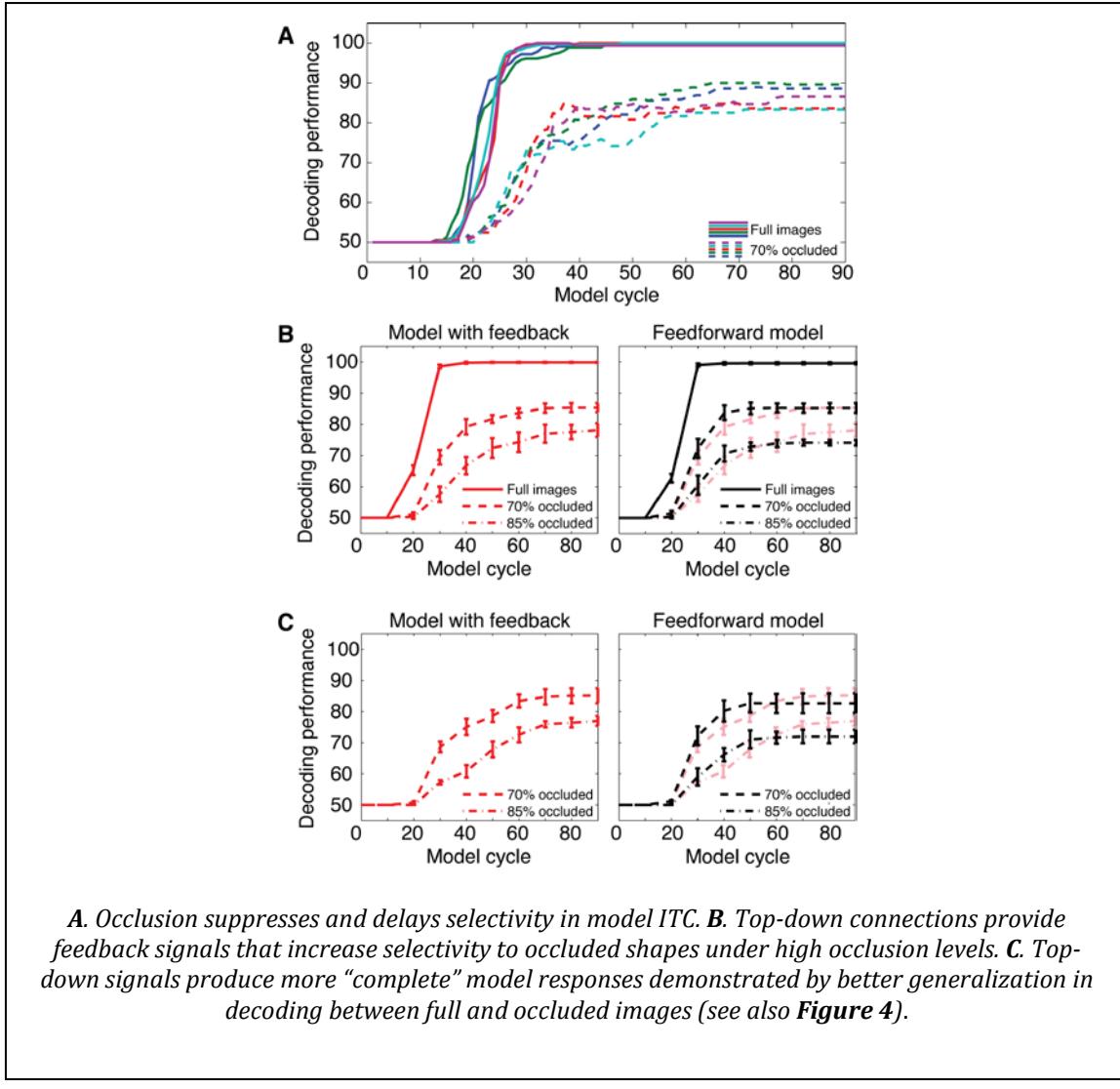


*Example runs of a model that includes both bottom-up and top-down connections at all layers. The connections are trained and the model iteratively cycles through bottom-up and top-down signals. The x-axis represents time in model cycles. A. The y-axis indicates the similarity (cosine distance) between the response vector at each layer for the occluded images and the unoccluded images. Each color represents a different layer in the model. From left to right, we increase the amount of unoccluded image from 50% to 70% to 90%. B. The y-axis indicates the similarity (cosine distance) between the response vector at each layer and at each cycle and the final response when the simulations ended.*

**Figure 5: Performance of the model for partial inputs (example)**

between the responses of the C2 units in the model and the labels corresponding to the 25 possible exemplars. The classifier works relatively well when a large fraction of the objects is visible (towards the right in **Figure 3 bottom**) and performance breaks down significantly when only a small fraction of the objects is visible

(towards the left in **Figure 3 bottom**). A first-order instantiation of top-down signals based on the schematic proposed in **Figure 3 top**, helps increase classification performance (blue line in **Figure 3 bottom**). While it does not completely recover performance, it can correspond to an enhancement of more than 10% in the amount of occlusion required to see the image. It should be noted that we do not expect performance to fully recover to the level in the near absence of occlusion. Even at the perceptual level, this is a very difficult task and recognizing an

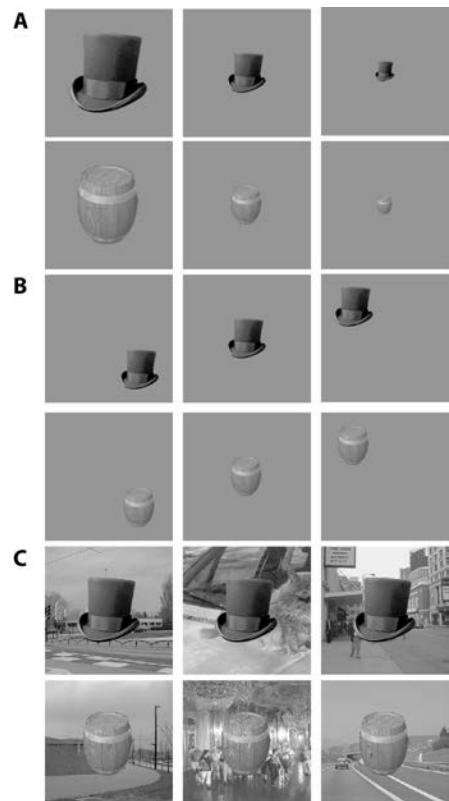


**Figure 6: Performance of the model for partial inputs (summary)**

object based on brief presentations of only a fraction of the information is extremely difficult.

The initial implementation described above included feedback signals only at the top of the hierarchical architecture. It is well established that top-down signals exist essentially at every level in neocortex (e.g. (Douglas and Martin, 2004)). We

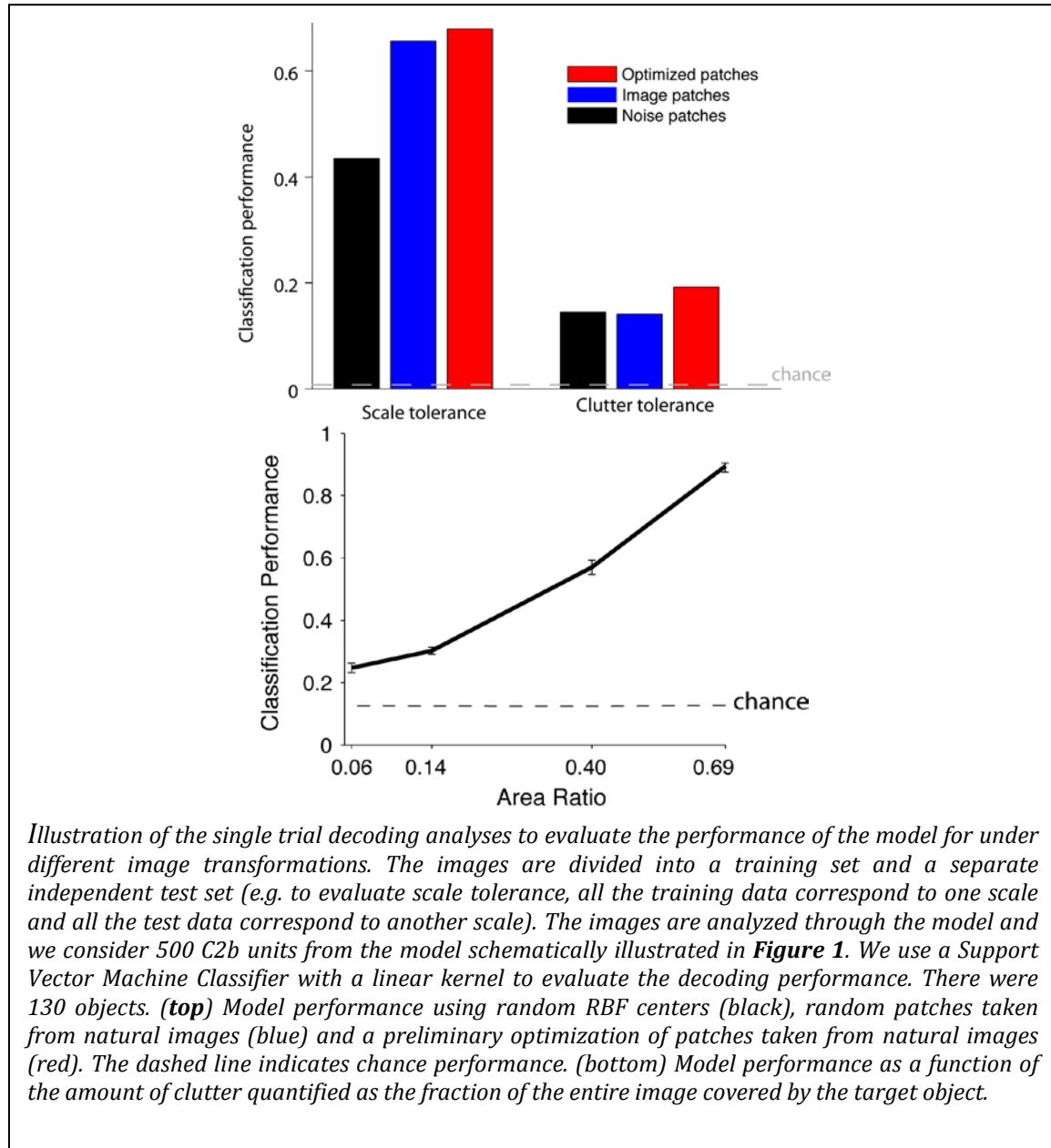
have begun to examine an architecture (O'Reilly et al., 2010) extending the work in (Serre et al., 2007) to include both bottom-up as well as top-down connections at every stage in the hierarchy (**Figure 4**). Computations throughout this circuitry occur in iterative cycles consisting of discrete steps of bottom-up and top-down activity. Example results are shown in **Figures 5-6**. In these simulations, we used the same set of 25 exemplar objects and bubbles technique from **Figures 3**. These simulations were performed on 4 out of 5 of the images from each category, using the last image for testing. The model was trained to categorize these 4 images (5-alternative forced choice). In this initial example training was done with only 60 invariance transformations per image. Effects would likely be larger with more training since the model would develop stronger features/sharper tuning. The preliminary simulations suggest that there is an increase in the similarity of the responses between occluded and unoccluded images with the number of bottom-up / top-down iterative cycles (**Figure 5A**). In particular, for the highest layers of the model (green and blue curves), the simulations show rather strong similarity suggesting that the unoccluded response is rescued through the iterative cycles. In the bottom plots (**Figure 5B**), we show the similarity between the responses at a given cycle and the final responses (when the simulations were stopped). These plots indicate the relative speed of convergence across different layers. The effects of occlusion include suppressing the amplitude as well as increasing the latency of the responses (**Figure 6A**). The top-down signals can play a role in enhancing these attenuated responses during occlusion (**Figure 6B**) and recovering a more complete representation of the object that better resembles the stored prototypes (**Figure 6C**). While these simulations and the classification performance results described in **Figure 4-6** are preliminary, they provide encouraging reason to suggest that the addition of top-down connections can significantly improve the model's capabilities to recognize objects in cluttered scenes or where objects are occluded.



*Illustration of the type of images used to characterize the performance of the model under a variety of image transformations. Only two main exemplar objects are shown here but the tests were conducted using 130 objects. A. Changes in object scale. B. Changes in object position. C. Objects inserted in natural images.*

**Figure 7: Example stimuli and transformations**

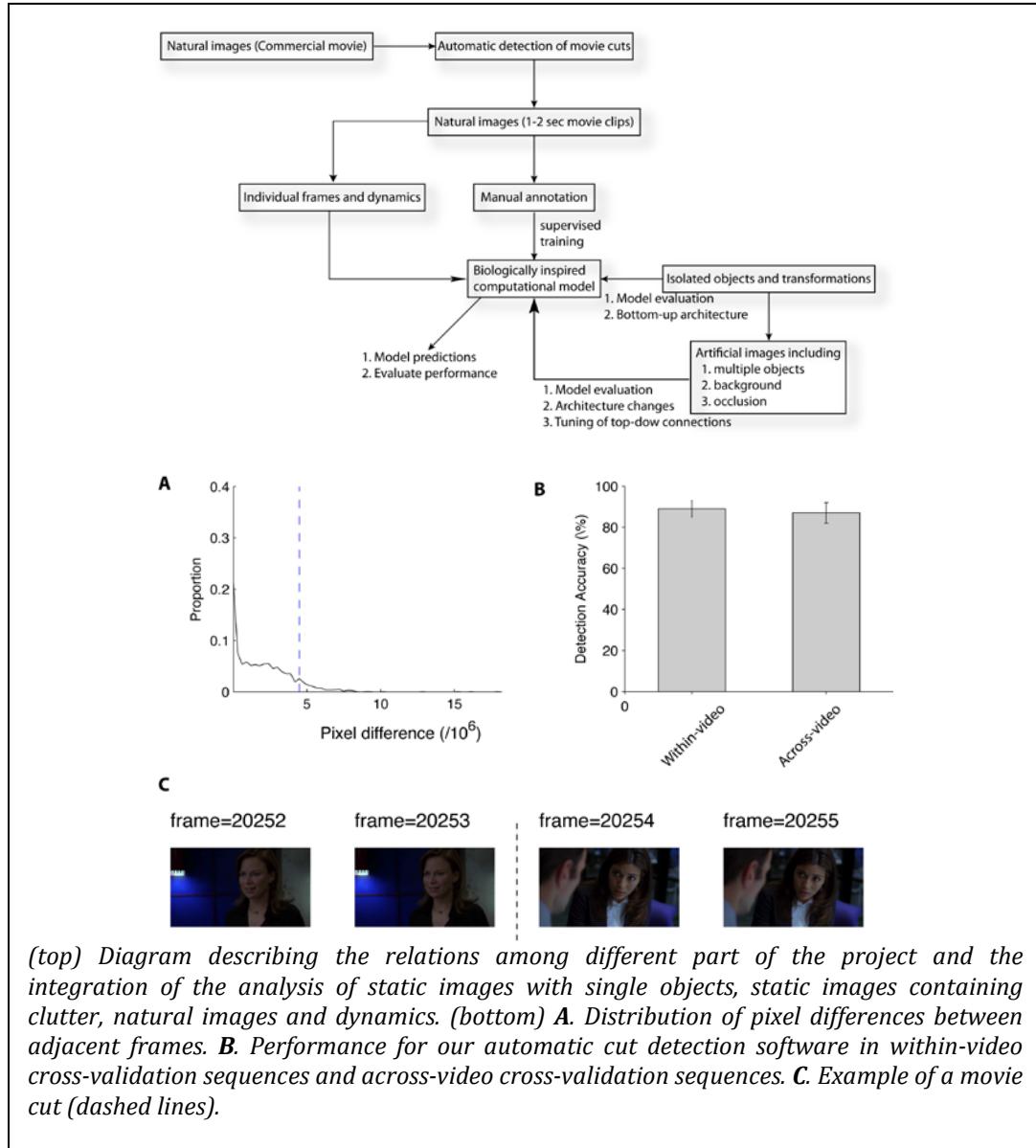
## 4.2 Characterizing the model's performance and RBF optimization



**Figure 8: Object transformations and model performance**

As schematically illustrated in **Figure 1**, one of the two key steps in the model involves the “tuning” operation whereby units along the hierarchy acquire increasing complexity in their feature preferences. The implementation in **Figure 1** uses Gaussian radial basis functions of the form: 
$$s_2 \mathbf{z}_j = \exp\left(-\frac{1}{2\sigma^2} \left\| c_1 \mathbf{z}_j - s_2 \mathbf{w}_j \right\|^2\right)$$
 where  $c_1 \mathbf{z}_j$  is the input from “C1 units” and the output  $s_2 \mathbf{z}_j$  of the “S2 units” depends on the centers defined in  $s_2 \mathbf{w}_j$ . Most of the results presented thus far

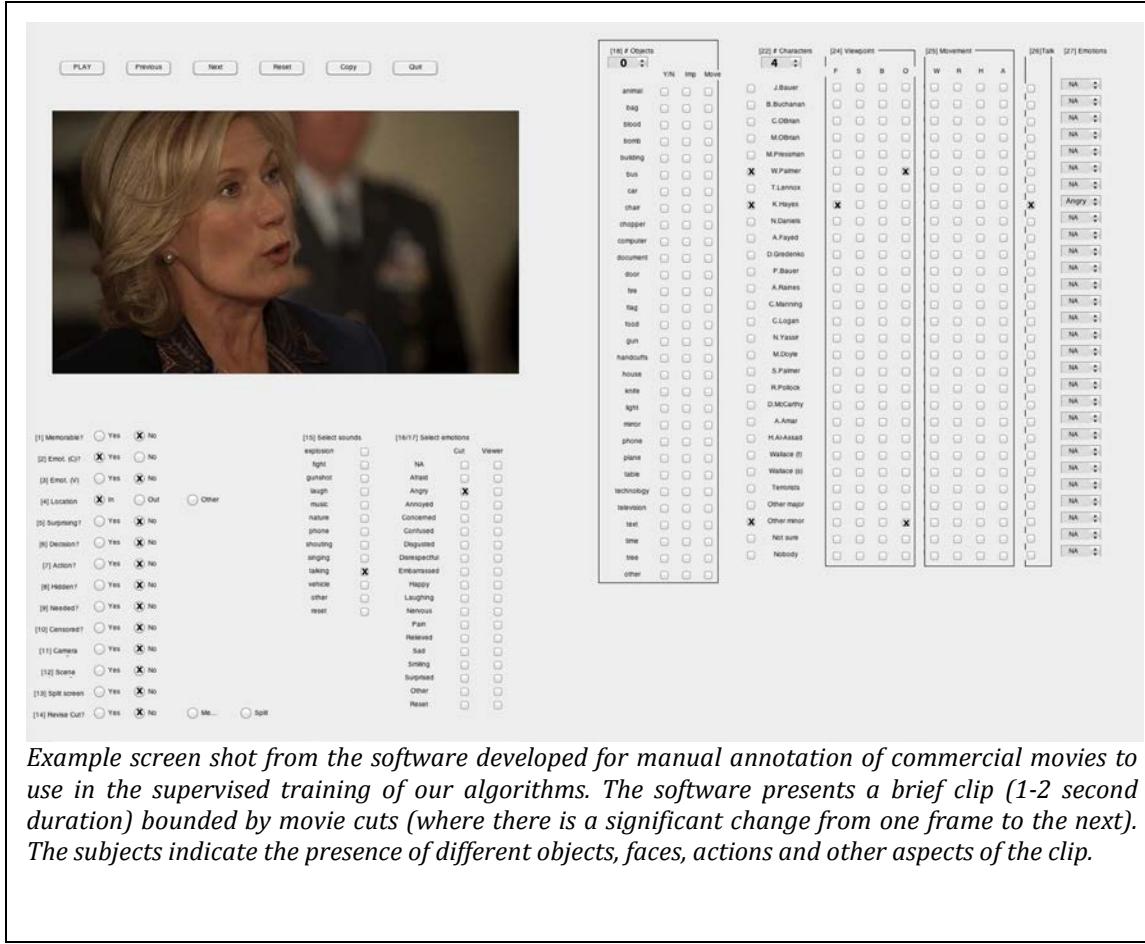
involve the use of centers defined by using arbitrary patches from natural images (e.g. Serre et al 2007). The use of “random defined” centers yields similar results. We have taken initial steps towards evaluating a genetic algorithm to select “good” RBF centers  $s_2 \mathbf{w}_j$  by considering a large library of possible centers and a cross-validation approach. This procedure was evaluated using a large library of images generated by considering 130 exemplar objects that changed in scale (**Figure 7A**), position (**Figure 7B**) or where objects were embedded in complex natural scenes (**Figure 7C**).



**Figure 9: Towards studying video sequences**

An initial proof-of-principle for the potential of optimizing the RBF centers is illustrated by the results in **Figure 8**. We use an SVM classifier with a linear kernel

to decode the identity of the image (among 130 possibilities) using a cross-validation approach. The red bars correspond to the use of “improved” RBF centers selected using a genetic algorithm. Although this simple optimization scheme was used only within a limited sample and only at the level of the S2 units in the model, we can already observe a small but significant improvement with respect to using random centers. This is particularly the case in the context of images with heavy clutter (**Figure 8 top**). Ultimately, performance in this type of recognition tasks will depend strongly on the amount of clutter (**Figure 8 bottom**).



Example screen shot from the software developed for manual annotation of commercial movies to use in the supervised training of our algorithms. The software presents a brief clip (1-2 second duration) bounded by movie cuts (where there is a significant change from one frame to the next). The subjects indicate the presence of different objects, faces, actions and other aspects of the clip.

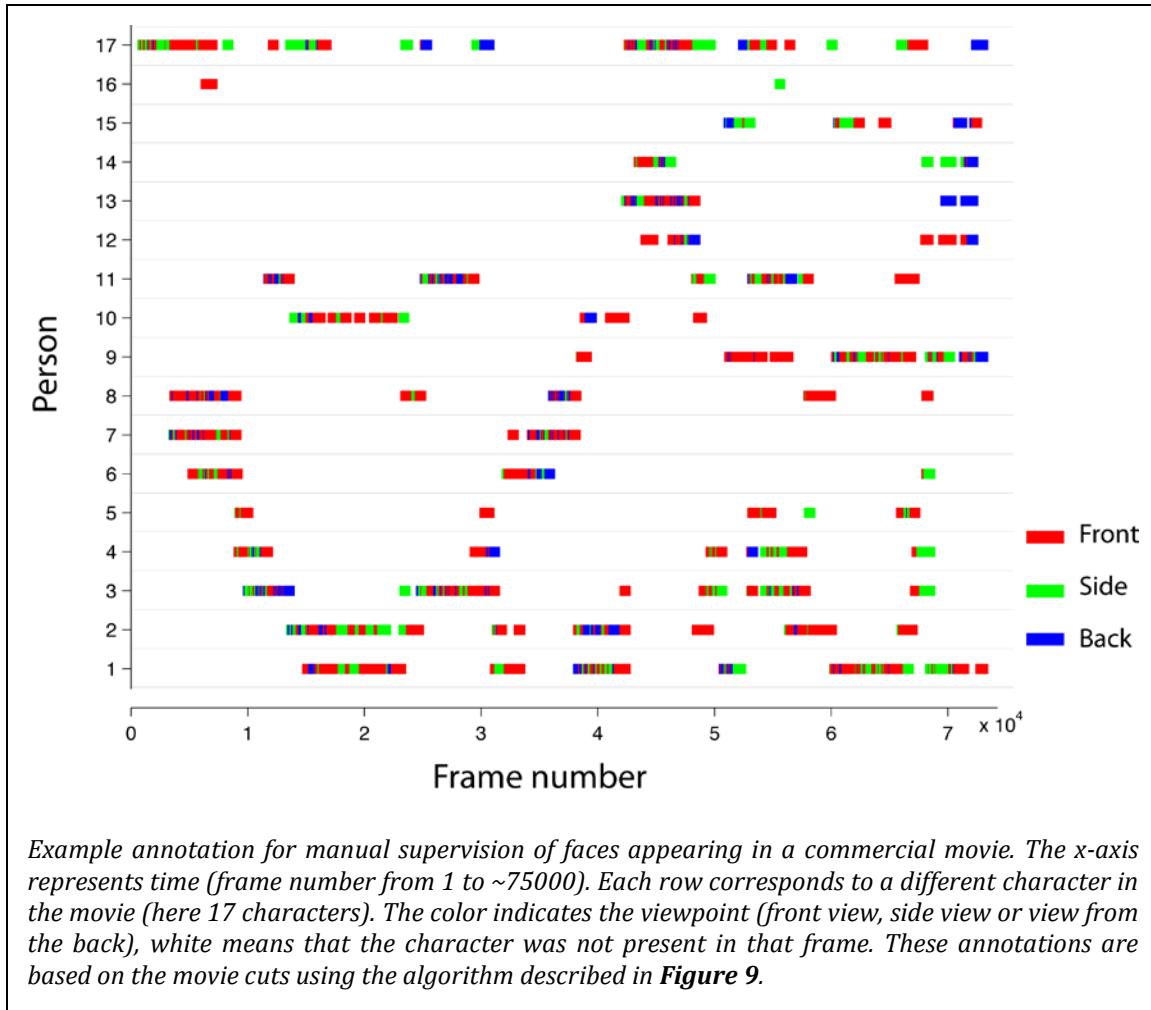
**Figure 10: Manual annotation scheme**

### 4.3 Recognition during dynamic viewing conditions

Most of our discussion so far has included static images. This is partly motivated by the large availability of static image data and by the observation that image recognition is quite rapid and efficient even from static information. Yet, image sequences (in particular the type of continuous image sequences obtained

from video cameras) offer the possibility of enhancing recognition by integrating information over time (see scheme in **Figure 9-11**).

Towards the goal of testing our models with dynamical information we have begun to prepare a database on annotated commercial video. The plan is to consider two commercial TV series because these series easily allow for partitioning data into training and test sets with internal controls. The initial step involves manual annotation as a basis for supervised training algorithms. The initial manual annotation will include basic characteristics of the overall scene (e.g. indoors versus outdoors), basic characteristics of the type of action involved (e.g. object movement versus dialogue), number and identity of the characters, emotions and other information that may be of potential interest for decoding through automatic vision.



**Figure 11: Example semi-supervised video annotations**

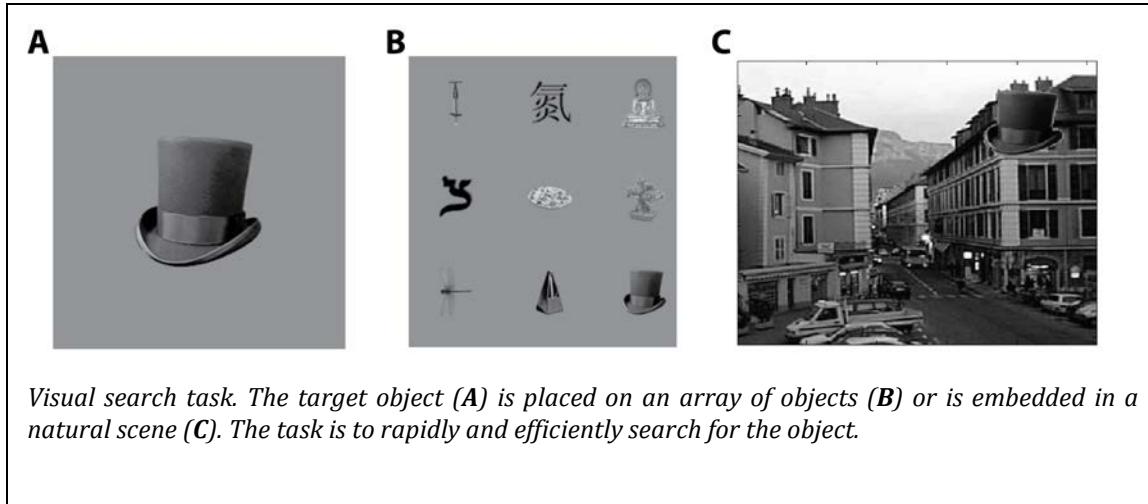
To begin with, we use as basic units the interval between two movie cuts. A movie cut is defined as a marked transition between two frames adjacent frames defined by a large distance in the pixel distribution (**Figure 9 bottom**). We have developed an automatic algorithm to detect these movie cuts automatically. The algorithm has a performance that is approximately 90% correct (the remaining 10%

typically consists of cuts that are hard to define such as instances where the camera moves continuously and spans a scene, or when the camera zooms in continuously on a given part of an image). These cuts are typically placed with a frequency of about 2 seconds. Using these intervals instead of annotating individual frames yields a savings of about 50-100x in the time required for annotation. Furthermore, these assignments of objects and faces to cuts will facilitate subsequent learning algorithms that exploit temporal correlations.

We have developed easy-to-use software for video annotation (**Figure 10**). The procedure starts with automatic detection of “cuts”, frames where the image content changes significantly from the previous one. We are in the process of manual annotation of two series, with 10 hours per series, a total of 20 hours of video data. The duration of each cut is approximately 1-2 seconds; this will yield several thousands of movie cuts, yielding a good size database for training our algorithms. An example of the output from this procedure is shown in **Figure 11**. Focusing here on faces, we provide the annotation output from one video sequence of 41-minute duration (73800 frames at 30 frames per second). Each indicates the presence and viewpoint for each one of 17 possible faces (white = absent, red = front view, green = side view, blue = view from the back). This is a video sequence taken from a commercial TV series (24) and includes complex backgrounds, multiple objects, changes in position, scale, illumination, etc. This provides a rich and complex database to evaluate the performance of our model in close to real-world scenarios.

The manual annotation will be used to supervise the training of our computational model for recognition. The analysis of this data set represents a considerable computational challenge. At 30 frames per second, 3600 second movies and approximately 20 movies, we are projecting the analysis of  $30 \times 3600 \times 20 \sim 2.2 \times 10^6$  frames. This serves as a good opportunity to enhance the speed of the computational model. We are working on rewriting parts of the code to improve speed and on evaluating how to adapt the code to incorporate the fact that large parts of the image may not change from one frame to the next (except for movie cuts).

**Figure 12** provides a glimpse of the initial steps towards evaluating the performance of the model in these video sequences. Our computational model will process each frame and temporal dependencies (e.g. correlations across frames) will be introduced into the model in the form of prior probabilities implemented through the feedback connections. These prior probabilities may significantly improve both speed as well as accuracy by taking advantage of the typically slow changes across adjacent frames.

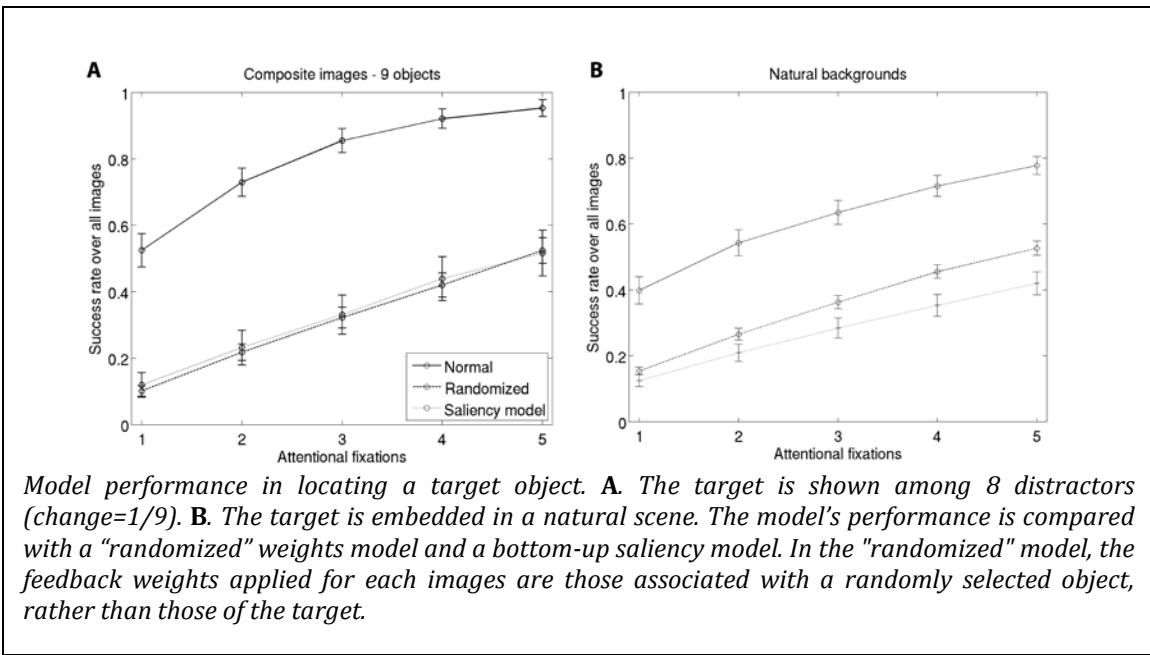


**Figure 12: Example stimuli for visual search task**

#### 4.4 Visual search

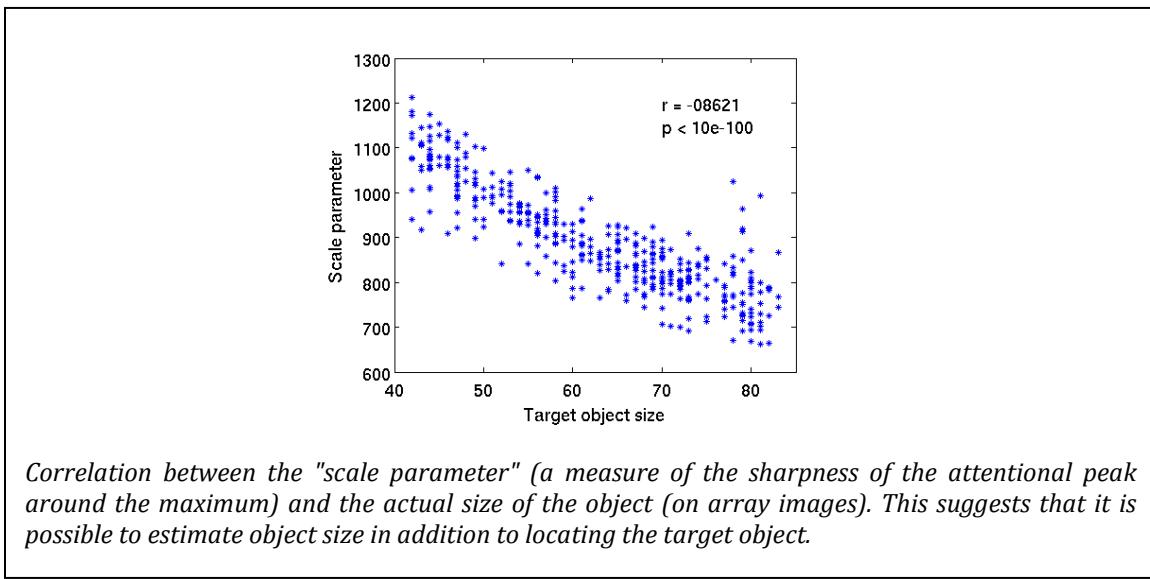
A central goal of this project, a key component of Aims 1-3, is the possibility of building biologically inspired computational algorithms capable of searching complex objects in cluttered scenes. During visual search, neural responses are modulated according to the features of the stimulus in the cell's receptive field (feature-based attention). Recent studies impose strong constraints on the mechanisms of this type of modulation. Here we describe a simple mechanistic model of feature-based attention for visual search. This model extends the previous work in bottom-up architectures (Serre et al 2007) and recent implementations of visual search processes (Chikkerur et al 2010).

Our model posits that a target-specific modulation is applied to a retinotopic area selective for visual features of intermediate complexity, with local normalization through divisive inhibition. This creates an "attentional map", where aggregate activity at any point tracks the match between local features and target features. The output of this map is then fed back to the visual system as a modulatory input, generating attentional effects. We propose a physiological interpretation in which the attentional map is computed by area LIP (which possesses the necessary characteristics for this computation) then fed back to FEF, and from there to V4. A schematic of the model is shown in **Figure 2**. We apply a computational implementation of this model to the task of finding target objects in either arrays of distractor objects arranged on a blank background, or in complex natural scenes (**Figure 12**). We show that the model significantly speeds up target localization when compared to either random fixations, or a pure saliency-based search. We demonstrate and explain the importance of normalization for successful matching between local input and target features.

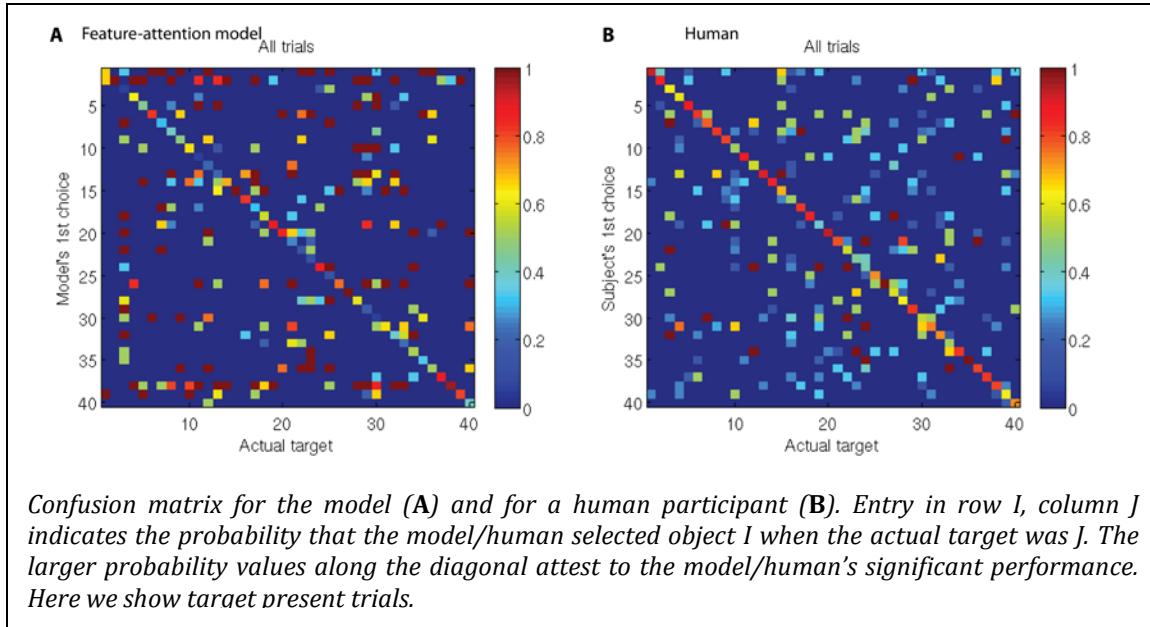


**Figure 13: The model can search for target objects**

We evaluate the performance of the model by considering the number of "fixations" required by the model to locate the target object. An example of the results is shown in **Figure 13A**, where a target is sought in an array of 9 objects (chance performance = 1/9). The feature attention model proposed here can detect the target object with >50% accuracy in the first fixation and significantly outperforms purely bottom-up approaches. Similar results are shown for objects embedded in natural backgrounds in **Figure 13B**.



**Figure 14: A normalization operation is critical for search model**



**Figure 15: Comparison between model and human performance**

In addition to locating the object, from the model's estimation of the extent of attentional modulation, it is possible to obtain an approximate estimation of the target object's size (**Figure 14**).

For comparison purposes, we are conducting psychophysics to evaluate human performance under the same conditions as the model. An example of these results is illustrated in the confusion matrices in **Figure 15** where we argue that both the model (A) and human participant (B) can correctly identify the target in most but not all trials.

#### 4.5 Improvements to prototypes

##### *Object completion and recognition of partially occluded objects.*

Initial tests that include training bottom-up and top-down connections yield significant improvement in recognition of partially occluded objects (**Figures 4-5**)

##### *Radial basis function optimization*

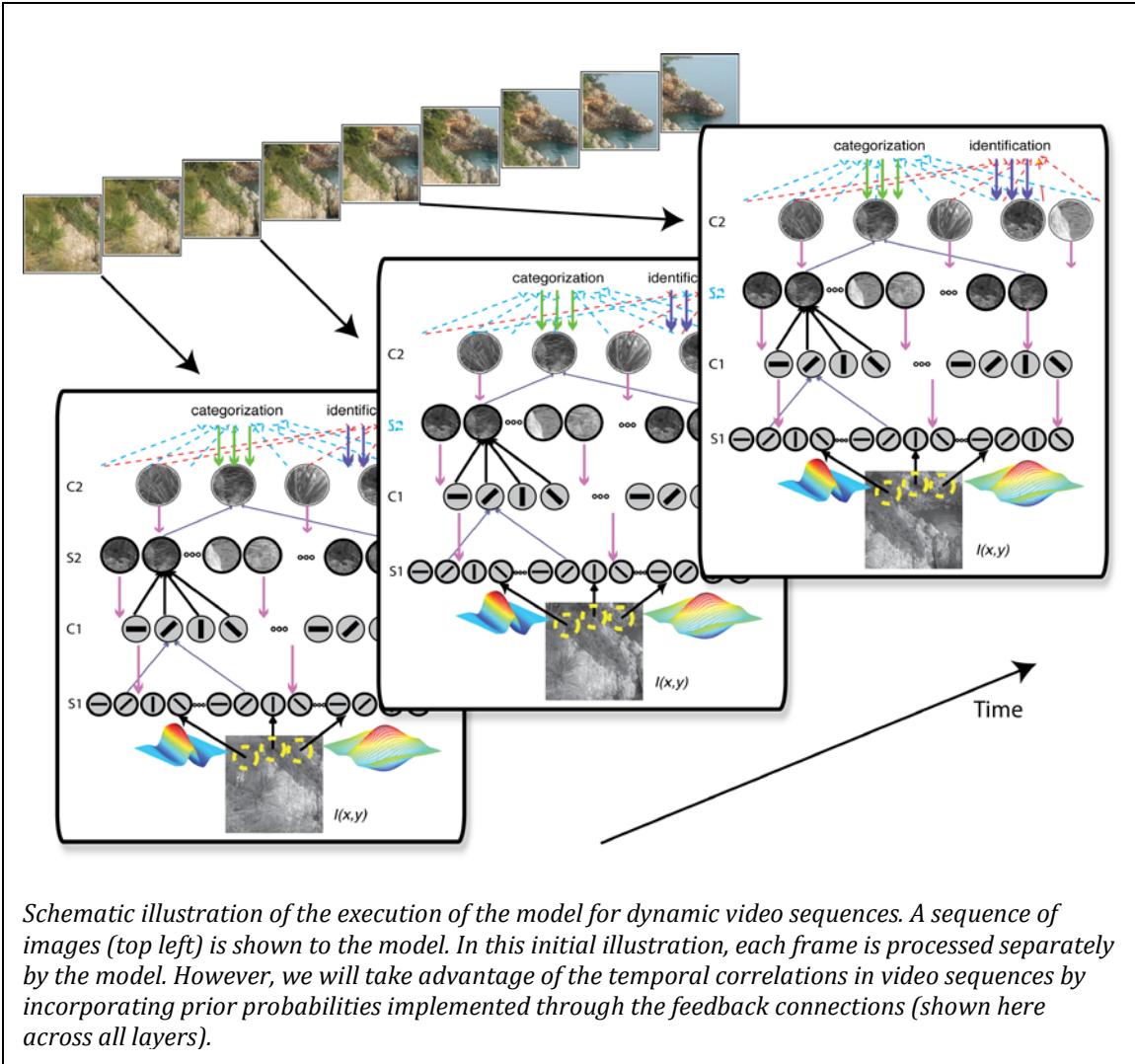
We have implemented an initial optimization step towards selecting the RBF centers at the S2 level in the model for improved performance (**Figures 6-7**).

##### *Automatic cut detection in video sequences*

We successfully developed an algorithm for automatic cut detection in video sequences (**Figure 8**)

##### *Semi-supervised annotation of video sequences*

We have made progress in the paradigm to annotate video sequences for supervision and training of our computational algorithm (**Figure 9-10**).



**Figure 16: Schematic proposal to study video sequences**

#### *Analysis of video sequences*

We have developed the code for the model that will be used for analyzing complex video sequences (**Figure 16**).

#### *Visual search in cluttered scenes*

We show in **Figures 11-14** that the feature attention model shows a remarkable performance in locating target objects in cluttered scenes.

#### *Comparison between human performance and machine performance in complex visual search tasks*

We show in **Figure 15** our initial comparison of performance in visual search between our model and human participants.

## 4.6 Publications

### *Journal publications*

Burbank K, Kreiman G (2012) Depression-biased reverse plasticity rule is required for stable learning at top-down connections. PLoS Computational Biology, 8(3):1-16.

Visual Population Codes: Toward a Common Multivariate Framework for Cell Recording and Functional Imaging. Edited by Nikolaus Kriegeskorte and Gabriel Kreiman. (2011). MIT Press. Cambridge, MA. ISBN-10:0-262-01624-9

Isik, L.\*, J.Z. Leibo\* and T. Poggio (2012). Learning and disrupting invariance in visual recognition with a temporal association rule. *Front. Comput. Neurosci.* 6:37.

Poggio, T. The Levels of Understanding framework, revised, MIT-CSAIL-TR-2012-014, CBCL-308, Massachusetts Institute of Technology, Cambridge, MA, May 31, 2012

Tan, C., J.Z. Leibo, and T. Poggio, Throwing Down the Visual Intelligence Gauntlet. Machine Learning for Computer Vision; eds Cipolla R., Battiato S., Farinella G.M., Springer: Studies in Computational Intelligence Vol. 411. July, 2012.

Chikkerur, S. and T. Poggio, Approximations in the HMAX Model, MIT-CSAIL-TR-2011-021/CBCL-298, Massachusetts Institute of Technology, Cambridge, MA, April 14, 2011

Leibo, J.Z., J. Mutch and T Poggio. How can cells in the anterior medial face patch be viewpoint invariant?, Presented at COSYNE 2011, Salt Lake City, UT. Available from Nature Precedings at [dx.doi.org/10.1038/npre.2011.5845.1](https://dx.doi.org/10.1038/npre.2011.5845.1)

Leibo, J.Z., J. Mutch and T Poggio, Learning to discount transformations as the computational goal of visual cortex, Presented at FGVC/CVPR 2011, Colorado Springs, CO. Available from Nature Precedings at [dx.doi.org/10.1038/npre.2011.6078.1](https://dx.doi.org/10.1038/npre.2011.6078.1)

Poggio, T. (sections with J. Mutch, J.Z. Leibo and L. Rosasco), The Computational Magic of the Ventral Stream: Towards a Theory, *Nature Precedings*, doi:10.1038/npre.2011.6117.1 July 16, 2011

### *Meetings and presentations*

Wyatte, D., Tang, H., Buia, C., Madsen, J., O'Reilly, R., & Kreiman, G. (2012). Object completion along the ventral visual stream: neural signatures and computational mechanisms. Computation and Systems Neuroscience (Cosyne) Annual Meeting.

Temporal constraints for visual object recognition: neurophysiological, behavioural and computational approaches. Kreiman. Bernstein Center for Computational Neuroscience. Berlin. Germany.

Collaborative research in computational neuroscience. June 2012. St. Louis. Missouri. Kreiman. Poggio.

Neural mechanisms and computational models for object recognition. Kreiman. University of Chicago. February 2012.

## 5. Conclusions

The development of biologically inspired computational algorithms for object recognition and search holds the promise to radically transform multiple domains of science and engineering including security applications, automatic search and analysis of web images and content and automatic navigation among others. Several convergent developments make progress towards these goals a realistic and attainable goal within the next years. These developments include the ever increasing amount and sophistication of computational power, our increased understanding of the mechanisms and circuits involved in recognition in the human brain, the availability of static and dynamic data sets to compare and test algorithms and the mathematics of learning.

As a proof of principle, this report highlights major progress towards addressing two of the most critical problems of real-world object recognition. The first one involves recognition from partial information. Under natural conditions, sensors typically acquire only partial information due to camouflage, perspective and occlusion. Reconstructing and identifying objects from partial information constitutes a formidable challenge. Purely bottom-up architectures only show limited power for object completion. Capitalizing on the theoretical understanding of recurrent networks and the power of feedback connections, here we show that an extension of bottom-up architectures that includes top-down signals can significantly improve performance in recognition of occluded objects.

A second domain that we illustrate in this report involves the central problem of identifying objects under heavy clutter. Searching for objects in cluttered scenes also constitutes a ubiquitous challenge in visual recognition. Here we provide evidence that top-down connections can be used to bias search towards the relevant features and thus significantly enhance visual search performance. Furthermore, we directly compare our feed-forward/feed-back model against human performance and we show that the model achieves a high degree of accuracy even when compared with humans in difficult rapid search tasks.

Finally, in an effort to further approximate real-world recognition problems, we provide initial steps towards the analysis of video sequences. This research endeavors to develop a semi-automatic annotation of video databases that can be

used to train and evaluate the performance of computational algorithms. We further describe initial steps on how to extend the current computational models to take advantage of dynamic information and the temporal correlations inherent in video sequences.

This is a golden age for the development of intelligent computer vision systems. The accuracy of such systems is still below human performance. Yet, major strides have been made during the last several years and there is enormous potential to come up with algorithms of direct applicability to a variety of problems in pattern recognition. These algorithms and their efficient implementation will radically transform multiple disciplines and lead to the development of commercial, military and clinical applications.

## 6. References

Bileschi S (2006) Street Scenes: towards scene understanding in still images. In: *Bran and Cognitive Science*. Cambridge: MIT.

Chikkerur S, Serre T, Poggio T (2009) A Bayesian inference theory of attention: neuroscience and algorithms. In: (MIT-CSAIL-TR, ed). Cambridge: MIT.

Chikkerur S, Serre T, Tan C, Poggio T (2010) What and where: A bayesian inference theory of attention. *Vision research* 50:2233-2247.

Chikkerur S, Serre T, Tan C, Poggio T (In Press) What and where: A bayesian inference theory of attention. *Vision Research*.

Douglas RJ, Martin KA (2004) Neuronal circuits of the neocortex. *Annu Rev Neurosci* 27:419-451.

Felleman DJ, Van Essen DC (1991) Distributed hierarchical processing in the primate cerebral cortex. *Cerebral cortex* 1:1-47.

Freedman D, Riesenhuber M, Poggio T, Miller E (2001) Categorical representation of visual stimuli in the primate prefrontal cortex. *Science* 291:312-316.

Fukushima K (1980) Neocognitron: a self organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics* 36:193-202.

Gosselin F, Schyns PG (2001) Bubbles: a technique to reveal the use of information in recognition tasks. *Vision research* 41:2261-2271.

Hubel D, Wiesel T (1959) Receptive fields of single neurons in the cat's striate cortex. *Journal of Physiology (London)* 148:574-591.

Hung C, Kreiman G, Poggio T, DiCarlo J (2005) Fast Read-out of Object Identity from Macaque Inferior Temporal Cortex. *Science* 310:863-866.

Mutch J, Knoblich U, Poggio T (2010) CNS: a GPU-based framework for simulating cortically-organized networks. In. Cambridge: MIT.

O'Reilly RC, Herd SA, Pauli WM (2010) Computational models of cognitive control. *Current opinion in neurobiology* 20:257-261.

Riesenhuber M, Poggio T (1999) Hierarchical models of object recognition in cortex. *Nature neuroscience* 2:1019-1025.

Serre T, Wolf L, Poggio T (2005a) Object Recognition with Features Inspired by Visual Cortex. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR). San Diego: IEEE Computer Society Press.

Serre T, Kouh M, Cadieu C, Knoblich U, Kreiman G, Poggio T (2005b) A theory of object recognition: computations and circuits in the feedforward path of the ventral stream in primate visual cortex. In, pp CBCL Paper #259/AI Memo #2005-2036. Boston: MIT.

Serre T, Kreiman G, Kouh M, Cadieu C, Knoblich U, Poggio T (2007) A quantitative theory of immediate visual recognition. *Progress In Brain Research* 165C:33-56.